

Analysis of a Population Genetics Model With Mutation, Selection, and Pleiotropy

S.N. Coppersmith,^{*†} Robert D. Blank,[‡] and Leo P. Kadanoff^{*}

February 1, 2008

Abstract

We investigate the behavior of a population genetics model introduced by (Waxman & Peck, 1998) incorporating mutation, selection, and pleiotropy. The population is infinite and continuous variation of genotype is allowed. Nonetheless, Waxman and Peck showed that if the degree of pleiotropy is large enough, in this model a nonzero fraction of the population can have identical alleles. This ‘condensed mode’ behavior appears in the limit of infinite times.

This paper explores the time-dependence of the distribution of alleles in this model. First, the model is analyzed using a recursion technique which enables the distribution of alleles to be calculated at finite times as well as in Waxman and Peck’s infinite-time limit. Second, both Waxman and Peck’s original model and a related model in which mutations occur continuously are mapped onto problems in quantum mechanics. In both cases, the long-time analysis for the biological model is equivalent to finding the nature of the eigenstates of the quantum problem. The condensed mode appears if and only if there is no bound state in the quantum problem.

We compare the behavior of the discrete- and continuous-time versions of the model. The results for the two cases are qualitatively similar, though there are some quantitative differences. We also discuss our attempts to correlate the statistics of DNA sequence variations with the degree of pleiotropy of various genes.

^{*}The James Franck Institute, The University of Chicago, 5640 S. Ellis Avenue, Chicago, IL 60637

[†]Corresponding author. Email address: snc@control.uchicago.edu; Fax number: (773) 702-5863.

[‡]The Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021

1 Introduction

Recently, (Waxman & Peck, 1998) introduced a simple population genetics model of an infinite population with a continuous distributions of alleles¹ incorporating pleiotropy (one gene affecting several characters of an organism). They demonstrated that when the number of characters affected is greater than two, the long-time steady state solution of their model can have a nonzero fraction of the population with identical alleles, and that this phenomenon does not occur if the number of characters affected per gene is two or fewer.

In this paper we consider their model and a simple variant of it, addressing several issues. The first three issues involve the mathematical analysis of the theory. First we study the time evolution of the distribution of alleles. In the model, the initial distribution of alleles is continuous, and at infinite time there is an infinitely narrow (δ -function) peak;² we ask how the distribution of alleles depends on time at long but finite times. Second, we describe how to construct time-dependent solutions by using an expansion in a sum of functions (Gaussians) particularly picked to meet the requirements of this problem. Third, we examine the relationship between the model with discrete generations used by Waxman and Peck (where mutations occur at discrete times, and fitness selection occurs continuously) and a continuous-time model in which both mutations and fitness selection occur continuously. The latter is a limiting case of the discrete model arising when the time between generations gets very short. We demonstrate that the qualitative behavior of the two models is the same, but that dependence of the behavior on the parameters can be different in the two models.

We show that the population genetics models can be mapped onto problems in quantum mechanics, specifically Bose-Einstein condensation³(de Groot *et al.*, 1950) and motion of a particle in a central potential.⁴ The mapping onto Bose condensation is performed on the discrete-time model in the limit that the fitness selection is very strong so that only organisms with fitnesses very near optimum survive each

¹Waxman and Peck's model is based on the continuum-of-alleles model of (Kimura, 1965). Other important investigations of pleiotropic models include (Wagner, 1989), (Lande, 1980), (Gavrilets & de Jong, 1993), and (Turelli, 1985).

²The Dirac delta-function, $\delta(t)$, is defined to be a function very sharply peaked around $t = 0$, but having a total weight $\int dt \delta(t)$ equal to one. One definition of such a function is the limit as z goes to positive infinity of $(2\pi z)^{-1/2} \exp[-t^2/(2z)]$. We shall deal with variables \vec{x} with N components. For such variables, $\delta(\vec{x})$ is defined to be a delta function in the first component of \vec{x} , times one in the second component, \dots . In this way, one has $\int d^N x \delta(\vec{x}) = 1$. (Spanier & Oldham, 1987)

³Bose-Einstein condensation in three dimensions is discussed in many texts on statistical mechanics, e.g., (Feynman, 1972).

⁴ See any book on quantum mechanics, e.g., (Schiff, 1968).

generation. The mapping onto the quantum particle in a central potential applies more generally. The simplest case has the potential independent of time. This case corresponds to the a limit of the population model in which fitness selection and mutations both occur continuously. The close relation between the continuous-time population biology models and Schrödinger’s equation has been exploited by Bürger⁵ to obtain many general results on the long-time behavior of models of the type studied here. In this paper, we focus on the mapping for a specific model, which yields a simple physical interpretation of the emergence of a unique genotype, and allows the extraction of the time-dependence of the fitness peak as well as comparison to the discrete-time model. We find that the behavior is qualitatively identical when the mutations are continuous and when they are discrete, though there are some quantitative differences.

We also discuss our attempts to relate the results from this model to the statistical properties of the sequences that are archived in various sequence databases. We analyze the statistical properties of gene sequences tabulated in genetic databases and attempt to correlate observed variations with estimates of the degree of pleiotropy of different genes in the databases. The results of these attempts are inconclusive.

This paper is organized as follows. In section 2 we introduce the theoretical models. Section 3 presents our analysis of the time evolution of the discrete version of the model, demonstrating that the behavior can be extracted analytically in a limit in which the typical jump caused by mutations is very large. In section 4 we present our analysis of the continuous-time model. Section 5 contains remarks on the relationship between the continuous- and discrete-time models. Section 6 presents our examination of DNA sequence archives and our attempts to correlate the statistics of documented sequence variations to the degree of pleiotropy of various genes. Section 7 recaps our main conclusions. The Appendix contains mathematical details of the analysis of the time-dependence of the continuous-time model aimed at establishing the long-time time-dependence.

2 The Models

One model we study is that of (Waxman & Peck, 1998). In this model, the population is infinite, the phenotypic variation is continuous, each gene affects N characters, and the effects of different genes are uncorrelated (no linkage disequilibrium or epistasis). The model assumes a very large population of haploid and asexual organisms with discrete generations. Let \vec{x} be a continuous vector with N components which represents characters which determine the viability of an organism, and $\phi(\vec{x}, t)$ be the

⁵A recent paper with references to previous work is (Bürger & Bomze, 1996).

normalized probability density that an organism in the population has the characters given by \vec{x} at time t . The normalization condition for the probability is given by an N -dimensional integral as

$$\int \phi(\vec{x}, t) d^N x = 1 . \quad (1)$$

At each generation an organism with characters \vec{x} survives viability selection with a probability proportional to $\exp(-x^2/2V_s)$. This Gaussian fitness selection will play an essential role in our solution of the model. Once each generation, a fraction θ of the population mutates; it is assumed that if a mutation occurs, then the probability that the mutant takes on the value \vec{x} given the parental value \vec{y} is $f(\vec{x} - \vec{y})$.

Since $\phi(\vec{x}, t)$ is a probability density, it is normalized at every time t . Thus, for this model, the equation for the time evolution of $\phi(\vec{x}, t)$ is:

$$\Gamma(t+1)\phi(\vec{x}, t+1) = (1-\theta)w_1(\vec{x})\phi(\vec{x}, t) + \theta \int f(\vec{x} - \vec{y})w_1(\vec{y})\phi(\vec{y}, t)d^N y , \quad (2)$$

where the fitness factor $w_1(\vec{x})$ is

$$w_1(\vec{x}) = \exp[-x^2/2V_s] . \quad (3)$$

The multiplicative factor, $\Gamma(t+1)$, in Eq. (2) ensures that the probability, $\phi(\vec{x}, t+1)$ is properly normalized at every time step. Integrating Eq. (2) over all \vec{x} gives the Waxman and Peck result for this normalization:

$$\Gamma(t+1) = \int \phi(\vec{x}, t)w_1(\vec{x})d^N x . \quad (4)$$

We follow Waxman and Peck in using a Gaussian function for the mutation probability:

$$f(\vec{x} - \vec{y}) = (2\pi m^2)^{-N/2} \exp[-(\vec{x} - \vec{y})^2/2m^2], \quad (5)$$

where m^2 describes the variance of the mutant effects for a single character.

In addition to this discrete-time model, we will consider a model in which both mutations and selection occur continuously in time. The relation between the discrete and continuous-time models can be seen explicitly by noting that the differential equation

$$\frac{\partial \phi(\vec{x}, t)}{\partial t} = \left[a(t) - \frac{x^2}{2V} \right] \phi(\vec{x}, t) \quad (6)$$

has the solution

$$\phi(\vec{x}, t + \tau) = \exp \left[\int_t^{t+\tau} ds \left(a(s) - \frac{x^2}{2V} \right) \right] \phi(\vec{x}, t). \quad (7)$$

Thus, model Eq. (2) can be written as a differential equation:

$$\begin{aligned} \frac{\partial \phi(\vec{x}, t)}{\partial t} = & \left[a(t) - \frac{x^2}{2V} \right] \phi(\vec{x}, t) \\ & + \theta \sum_n \delta(t - n\tau) \int d\vec{y} f(\vec{x} - \vec{y}) [\phi(\vec{y}, t) - \phi(\vec{x}, t)] , \end{aligned} \quad (8)$$

where τ is the interval between mutations. Here, $a(t)$ is determined by the normalization requirement of Eq. (1) and $\delta(t)$ is the Dirac delta function. In dynamical systems theory, equations with a sum of time-delta functions are said to be ‘kicked’ (Lichtenberg & Liberman, 1992). When the interval between kicks (here τ) goes to zero, the sum of delta-function is replaced by its time average, so that Eq. (8) reduces to:

$$\frac{\partial \phi(\vec{x}, t)}{\partial t} = \left[a(t) - \frac{x^2}{2V} \right] \phi(\vec{x}, t) + \Theta \int d\vec{y} f(\vec{x} - \vec{y}) [\phi(\vec{y}, t) - \phi(\vec{x}, t)] , \quad (9)$$

where Θ is the rate of mutations per unit time, θ/τ , and $a(t)$ is determined by the normalization requirement, Eq. (1.)

In the next few sections we calculate the evolution of $\phi(\vec{x}, t)$ for different N , with an emphasis on the behavior near $\vec{x} = 0$ at long times. In section 3 we examine the discrete-time model Eq. (2); section 4 discusses the continuous time equation Eq. (9). In section 5 we discuss the relationship between the results for the continuous and discrete-time equations.

3 Time Evolution of Discrete Equations

In this section we consider the discrete-time evolution Eq. (2). We will solve it by exploiting the following observation. Suppose at some time t $\phi(\vec{x}, t)$ is a normalized Gaussian with variance α , so that $\phi(\vec{x}, t) = G_\alpha(\vec{x})$, where⁶

$$G_\alpha(\vec{x}) = (2\pi\alpha)^{-N/2} \exp \left[-\frac{x^2}{2\alpha} \right] . \quad (10)$$

Eq. (2) then implies that $\phi(\vec{x}, t+1)$ is the sum of two Gaussians. Indeed, at any finite time t , $\phi(\vec{x}, t)$ is the sum of finitely many Gaussians.

To see how this works, notice that Eq. (2) involves the processes of multiplication and of convolution with a Gaussian. If one multiplies two Gaussians, one gets another Gaussian according to the rule:

$$G_\alpha(\vec{x}) G_\beta(\vec{x}) = \left(\frac{\gamma}{2\pi\alpha\beta} \right)^{N/2} G_\gamma(\vec{x}), \quad \text{where } 1/\gamma = 1/\alpha + 1/\beta . \quad (11)$$

⁶Strictly speaking, α is the variance of one component of \vec{x} .

Convolution has slightly different rules. A textbook integration yields the simple result:

$$\int G_\alpha(\vec{x} - \vec{y}) G_\beta(\vec{y}) d^N y = G_\delta(\vec{x}) , \quad \text{where } \delta = \alpha + \beta . \quad (12)$$

Instead of doing integrals at each time step, we can write recursion relations for the evolution of the widths and amplitudes of the Gaussians. We then solve these equations in the limit $V_s \ll m^2$. Our method of solution has similarities with that used by (Kingman, 1978) on a simpler model.

To perform the calculation explicitly, first consider a warm-up problem—the case of no mutations ($\theta = 0$), which is described by the equation:

$$\Gamma(t+1)\phi(\vec{x}, t+1) = \exp[-x^2/2V_s]\phi(\vec{x}, t) , \quad (13)$$

where $\Gamma(t+1)$ is defined in Eq. (4). We assume that, at time t , $\phi(\vec{x}, t)$ is a normalized Gaussian with variance $\alpha(t)$:

$$\phi(\vec{x}, t) = G_{\alpha(t)}(\vec{x}) \quad (14)$$

The multiplication process of Eq. (11) implies that the value of $\phi(\vec{x}, t+1)$ is, once again a Gaussian of the form $G_{\alpha(t+1)}(\vec{x})$, but that the new value of α is given by $1/\alpha(t+1) = 1/\alpha(t) + 1/V_s$. This equation for $\alpha(t)$ has the solution:

$$\alpha(t) = \frac{\alpha_0}{1 + \alpha_0 t/V_s} , \quad (15)$$

where $\alpha_0 = \alpha(t=0)$. Thus the population distribution remains Gaussian, and at long times, its width scales as $|\vec{x}| \sim t^{-1/2}$. As expected, fitness selection continually narrows the distribution.

Now we examine the full Eq. (2) with $\theta > 0$. We first calculate $\phi(\vec{x}, t=1)$, given that $\phi(\vec{x}, t=0)$ is a Gaussian. There are two terms in the equation. Each involves the multiplication of Gaussians, and the mutation term also involves a convolution. By the rules of Eqs. (11) and (12), $\phi(\vec{x}, t=1)$ is the sum of two Gaussians of the form:

$$\phi(\vec{x}, 1) = (1 - \theta)G_{\beta_1(\alpha)}(\vec{x}) + \theta G_{\beta_2(\alpha)}(\vec{x}), \quad (16)$$

where the two β 's are given by:

$$\beta_1(\alpha) = \frac{\alpha}{1 + \alpha/V_s} , \quad (17)$$

$$\beta_2(\alpha) = m^2 + \beta_1(\alpha) , \quad (18)$$

It follows immediately that if $\phi(\vec{x}, t)$ is the sum of finitely many Gaussians, then $\phi(\vec{x}, t+1)$ is also the sum of finitely many Gaussians. Explicitly, we write $\phi(\vec{x}, t)$ in the form:

$$\phi(\vec{x}, t) = \sum_i A_i(t) G_{\alpha_i(t)}(\vec{x}) , \quad (19)$$

with the normalization condition that $\sum_i A_i = 1$. Then $\phi(\vec{x}, t + 1)$ can be written:

$$\Gamma(t + 1)\phi(\vec{x}, t + 1) = (1 - \theta) \sum_i B_i(t) G_{\beta_1(\alpha_i(t))}(\vec{x}) + \theta \sum_i B_i(t) G_{\beta_2(\alpha_i(t))}(\vec{x}) , \quad (20)$$

with:

$$B_i(t) = A_i(t) [1 + \alpha_i(t)/V_s]^{-N/2} , \quad (21)$$

$$\Gamma(t + 1) = \sum_i B_i(t) . \quad (22)$$

Eqs. (20-22) simplify considerably when $V_s/m^2 \ll 1$. In this case, the characteristic width of the mutation probability function, f , is much greater than the width produced by one step of the fitness narrowing. In this big-mutation-jump limit, $\beta_1(m^2) \approx V_s$, $\beta_2(m^2) \approx m^2$, $\beta_1(V_s/n) = V_s/(n + 1)$, $\beta_2(V_s/n) \approx m^2$.⁷ (Here, n is any positive integer.)

Consider the time evolution of an initial distribution consisting of a Gaussian with variance $\alpha = m^2$. After one time step, the distribution splits into two Gaussians, an unmutated population with variance $\beta \approx V_s$ and a mutated population with variance $\beta \approx m^2$. After two time steps, the population consists of three Gaussians with β values $\approx m^2$, V_s , and $V_s/2$. After n steps, the population consists of $n + 1$ Gaussians with β values m^2 , V_s , $V_s/2$, ..., V_s/n . We define $\alpha_0 = m^2$, $\alpha_i = V_s/i$ for $i \geq 1$. (Note that these α 's are time-independent.) If we start with an initial distribution which is a Gaussian of width much greater than m , then after n steps, the distribution consists of $n + 1$ Gaussians with β values α_0 , ..., α_n . We describe the time evolution using rate equations for the populations in these states. Fig. 1 is a sketch of the transitions that can occur between the different α 's. Note that at a given time, a state α_i makes transitions to two states, α_0 and α_{i+1} .

We write $\phi(\vec{x}, t)$ as:

$$\phi(\vec{x}, t) = \sum_{i=0}^t A_i(t) G_{\alpha_i}(\vec{x}) , \quad (23)$$

where the $A_i(t)$ must obey:

$$A_0(t > 0) = \theta \quad (24)$$

$$A_1(t > 0) = \frac{(1 - \theta)}{\Gamma(t)} Q A_0(t - 1) , \quad (25)$$

$$A_i(t) = \frac{1 - \theta}{\Gamma(t)} \left[\frac{i - 1}{i} \right]^{N/2} A_{i-1}(t - 1) \quad (i > 1, t > 0) , \quad (26)$$

⁷Because the newly-mutated population is a Gaussian of width determined by m^2 for any distribution, this approximation is equivalent to the ‘‘house-of-cards’’ approximation introduced by (Kingman, 1978) and investigated in (Turelli, 1984), (Barton & Turelli, 1989). One can show that the results derived in this section are robust when this approximation is not made, so long as one is the regime where $V_s \ll m^2$.

and

$$\Gamma(t+1) = QA_0(t) + \sum_{i=1}^t A_i(t) \left(\frac{i}{i+1} \right)^{N/2}. \quad (27)$$

We have defined $Q = (1 + m^2/V_s)^{-N/2}$; note that $Q \ll 1$ in the limit we consider.

First consider the first two steps of the evolution. Starting with the initial condition $A_0(t=0) = 1$, $A_i(t=0) = 0$, one finds:

$$\begin{aligned} A_0(t=1) &= \theta, \\ A_1(t=1) &= 1 - \theta. \end{aligned} \quad (28)$$

So far not much has happened. But a key thing happens at the next step:

$$\begin{aligned} A_0(t=2) &= \theta, \\ A_1(t=2) &= \frac{\theta Q(1-\theta)}{\theta Q + (1-\theta)2^{-N/2}}, \\ A_2(t=2) &= \frac{(1-\theta)^2 2^{-N/2}}{\theta Q + (1-\theta)2^{-N/2}}. \end{aligned} \quad (29)$$

Note that $A_1(t=2) \propto Q \ll 1$. In fact, for any $t > 2$, all A_i 's with $0 < i < t$ are proportional to Q . However, as $t \rightarrow \infty$, the number of these terms diverges. So it is not obvious whether as $t \rightarrow \infty$ a solution exists where $A_0(t) = \theta$, $A_t(t) = O(1)$, and all other $A_i(t)$'s are small, which would imply that the Gaussian describing the unmutated population contains a nonzero fraction of the total population. We will find that such a solution can exist only when $N > 2$.

First we find the steady-state solutions in the long-time limit $t \rightarrow \infty$. In steady state, the time arguments on the A_i and on Γ can be dropped, and the recursion relations become:

$$A_0 = \theta \quad (30)$$

$$A_1 = \theta Q v, \quad (31)$$

$$A_i = v \left[\frac{i-1}{i} \right]^{N/2} A_{i-1} \quad (i > 1), \quad (32)$$

and

$$\frac{1-\theta}{v} = \theta Q + \lim_{t \rightarrow \infty} \sum_{i=1}^t A_i \left(\frac{i}{i+1} \right)^{N/2}, \quad (33)$$

where we have defined $v = (1-\theta)/\Gamma$. We explicitly allow for the possibility that the amplitude $A_\infty \equiv \lim_{t \rightarrow \infty} A_t(t)$ is nonzero.

The solution to Eqs. (30-33) is:

$$\begin{aligned} A_0 &= \theta, \\ A_i &= \theta Q i^{-N/2} v^i, \quad i \geq 1, \end{aligned} \quad (34)$$

where v must satisfy:

$$\frac{1-\theta}{v} = A_\infty + \theta Q \sum_{i=0}^{\infty} v^i \left(\frac{1}{i+1} \right)^{N/2}. \quad (35)$$

Note that $v \leq 1$, for otherwise the A_i cannot sum to unity, which is required for normalization of the probability.⁸ Since as v increases the left hand side of Eq. (35) monotonically decreases and the right hand side monotonically increases, a solution with $A_\infty = 0$ can exist only if $\theta Q \sum_{i=0}^{\infty} \left(\frac{1}{i+1} \right)^{N/2} > 1 - \theta$. Conversely, if this inequality is not satisfied, then we expect $A_\infty > 0$. Since $A_\infty \sim \lim_{t \rightarrow \infty} v^t$, we require $v = 1$ if A_∞ is nonzero. We show below that such a solution is the long-time limit of the solution of the time-dependent equations.

First consider the case when $A_\infty \neq 0$. Since the sum in Eq. (35) converges if $N > 2$, and diverges otherwise, we see that $A_\infty \neq 0$ can occur only if $N > 2$. When the sum does converge, the unmutated fitness peak, whose amplitude is A_∞ and whose width is given by Eq. (15), can contain a nonzero fraction of the total weight. Evaluating Eq. (33) with $v = 1$, we find $A_\infty = 1 - \theta - \theta Q \zeta(N/2)$, where $\zeta(N/2)$ is the Riemann zeta function (Carrier *et al.*, 1983; Abramowitz & Stegun, 1972).

Now we consider the possibility of a solution in which the unmutated population constitutes an infinitesimal fraction of the total population as $t \rightarrow \infty$. Since $A_\infty = 0$, we must have:

$$\frac{1-\theta}{\theta Q} = \Phi(N/2, v), \quad (36)$$

where $v < 1$ and $\Phi(a, x)$ is the polylogarithm function (Lewin, 1981):

$$\Phi(a, x) \equiv \sum_{i=1}^{\infty} \frac{1}{i^a} x^i. \quad (37)$$

When $a > 1$, $\Phi(a, x = 1) = \zeta(a)$ is finite, where again $\zeta(z)$ is the Riemann zeta function. Therefore, for $N > 2$, the right hand side of Eq. (36) is bounded as $v \rightarrow 1$, and a solution exists only if $(1 - \theta)/(\theta Q) < \zeta(N/2)$. In the $Q \ll 1$ limit that we have assumed, this happens only when $1 - \theta$ is very small also. If this condition is not satisfied, then a nonzero fraction of the population must be in the unmutated state.

When $N = 2$, $\Phi(1, v) = -\log(1 - v)$, and in the limit $\theta Q/(1 - \theta) \ll 1$, one finds

$$v \approx 1 - \exp \left[-\frac{1 - \theta}{\theta Q} \right], \quad (38)$$

⁸This inequality is derived in (Waxman & Peck, 1998), and is closely related to results derived by Bürger and collaborators (Bürger, 1986), (Bürger, 1988), (Bürger & Hofbauer, 1994), (Bürger & Bomze, 1996)

so that:

$$\Gamma \approx (1 - \theta) \left(1 + \exp \left[-\frac{1 - \theta}{\theta Q} \right] \right) \quad (N = 2). \quad (39)$$

When $N = 1$, the leading order term as $v \rightarrow 1$ can be calculated by noting that, for all $v < 1$, $\Phi(\frac{1}{2}, v)$ obeys the bounds:⁹

$$\int_1^\infty dx \frac{1}{\sqrt{x}} v^x < \Phi(\frac{1}{2}, v) < \int_0^\infty dx \frac{1}{\sqrt{x}} v^x. \quad (40)$$

The integrals both converge at their lower limits. They also converge at their upper limits when $v < 1$, but not when $v = 1$. The behavior near the upper limit can be used to estimate that, as v approaches 1,

$$\begin{aligned} \Phi(\frac{1}{2}, v) &\sim \int_0^\infty dx \frac{1}{\sqrt{x}} \exp[-x|\ln v|] \\ &= \sqrt{\frac{\pi}{|\ln v|}} \int_0^\infty dt \exp[-t^2] \\ &= \frac{\pi}{2\sqrt{|\ln v|}} \\ &\approx \pi/(2\sqrt{1-v}). \end{aligned} \quad (41)$$

Thus, $v \approx 1 - [(\pi\theta Q)/(2(1 - \theta))]^2$, and:

$$\Gamma \approx (1 - \theta) \left(1 + \left(\frac{\pi\theta Q}{2(1 - \theta)} \right)^2 \right) \quad (N = 1). \quad (42)$$

Now we calculate the function $\phi(\vec{x}, t \rightarrow \infty)$.¹⁰ First we consider the case when ϕ has a δ -function piece.

In this regime, as $t \rightarrow \infty$, $A_0 = \theta$, $A_i = \theta Q(1/i)^{N/2}$ ($i \geq 1$), and A_∞ , the weight in the δ -function, is $A_\infty = 1 - \theta - \theta Q\zeta(N/2)$, where $\zeta(N/2)$ is the Riemann zeta

⁹These bounds follow because the integrand is monotonically decreasing. For any positive integer j and $v \leq 1$, one has $\int_j^{j+1} dx \frac{1}{\sqrt{x}} v^x < \frac{1}{\sqrt{j}} v^j < \int_{j-1}^j dx \frac{1}{\sqrt{x}} v^x$.

¹⁰The form of the distribution calculated by Waxman and Peck (their footnote 29) does not apply here because they require simultaneous validity of the inequalities $\theta V_s/m^2 \ll 1$, $m^2/V_s \ll 1$, and $N \ll V_s/m^2$. Our calculation applies whenever $V_s/m^2 \ll 1$.

function. Therefore,

$$\begin{aligned}
\phi(\vec{x}, t \rightarrow \infty) &= (1 - \theta - \theta Q \zeta(N/2)) \delta(x) + \theta (2\pi m^2)^{-N/2} \exp \left[-\frac{x^2}{2m^2} \right] \\
&\quad + \theta Q \sum_{n=1}^{\infty} (1/n)^{N/2} (2\pi V_s/n)^{-N/2} \exp \left[-\frac{nx^2}{2V_s} \right] \\
&= (1 - \theta - \theta Q \zeta(N/2)) \delta(x) + \theta (2\pi m^2)^{-N/2} \exp \left[-\frac{x^2}{2m^2} \right] \\
&\quad + \theta Q (2\pi V_s)^{-N/2} \left(\exp \left[\frac{x^2}{2V_s} \right] - 1 \right)^{-1}. \tag{43}
\end{aligned}$$

Thus we see that when $\phi(\vec{x}, t \rightarrow \infty)$ has a δ -function piece, there is in addition a divergent contribution at small x , proportional to $1/x^2$. When the distribution is smooth, we have $A_0 = \theta$ and $A_i = \theta Q (1/i)^{N/2} v^i$ ($i \geq 1$), and $v < 1$, and we obtain:

$$\begin{aligned}
\phi(\vec{x}, t \rightarrow \infty) &= \theta (2\pi m^2)^{-N/2} \exp \left[-\frac{x^2}{2m^2} \right] \\
&\quad + \theta Q (2\pi V_s)^{-N/2} \left(v^{-1} \exp \left[\frac{x^2}{2V_s} \right] - 1 \right)^{-1}. \tag{44}
\end{aligned}$$

The sums that arise here are identical to those that come up in the calculation of Bose-Einstein condensation for an ideal Bose gas (de Groot *et al.*, 1950). The weight in the δ -function in the genotype distribution, A_∞ , is analogous to the condensate fraction in the Bose condensation problem. The parameter N , which here describes the number of traits affected by a mutation, is the number of spatial dimensions in the Bose gas calculation. That the superfluid fraction must be zero for a Bose gas in one and two dimensions is a special case of a general result in the theory of phase transitions (Mermin & Wagner, 1966), (Hohenberg, 1967), (Frölich & Pfister, 1981).

We now discuss the time evolution of the population distribution. We begin with some qualitative remarks. When $v < 1$, the sum in the solution Eq. (35) converges geometrically. Therefore, in this regime one expects the approach to the $t \rightarrow \infty$ limit to be exponential. In the regime where a δ -function contribution is present as $t \rightarrow \infty$, the δ -function is the long-time limit of a Gaussian describing the unmutated population. The squared width of this Gaussian narrows as $1/t$, so we expect the approach to the long-time limit to be power-law. These expectations are supported by the explicit calculation that we now present. We, in fact, show that the long-time corrections to the amplitude of the δ -function are also proportional to $1/t$.

We again write recursion relations describing the transitions between the various Gaussians, now allowing for explicit time dependence in the $A_i(t)$. Defining $v(t) =$

$(1 - \theta)/\Gamma(t + 1)$, these recursion relations are, for $t > 0$:

$$A_0(t) = \theta , \quad (45)$$

$$A_1(t) = Qv(t - 1)A_0(t - 1) , \quad (46)$$

$$A_i(t) = v(t - 1) \left(\frac{i - 1}{i} \right)^{N/2} A_{i-1}(t - 1) , \quad (47)$$

$$\frac{1 - \theta}{v(t)} = QA_0(t) + \sum_{i=1}^t A_i(t) \left(\frac{i}{i + 1} \right)^{N/2} . \quad (48)$$

For the initial conditions $A_0(t = 0) = 1$ and $A_{i>0}(t = 0) = 0$, these recursion relations have the solution:

$$A_i(t) = Q \prod_{t'=t-i}^{t-1} v(t') \left(\frac{1}{i} \right)^{N/2} A_0(t - i + 1) . \quad (49)$$

The self-consistency condition is:

$$\frac{1 - \theta}{v(t)} = \frac{Q}{v(t)} \left[\sum_{j=1}^{t+1} \left(\frac{1}{j} \right)^{N/2} A_0(t - j) \prod_{t'=t-j+1}^t v(t') \right] . \quad (50)$$

Now we separate out explicitly the term $j = t + 1$; this is reasonable because $v(t' = 0)$ is larger than all the other v 's by a factor proportional to $1/Q$, yielding:

$$\frac{1 - \theta}{v(t)} = \frac{Q\theta}{v(t)} \sum_{j=1}^t \left(\frac{1}{j} \right)^{N/2} \prod_{t'=t-j+1}^t v(t') + (1 - \theta) \left(\frac{1}{t + 1} \right)^{N/2} \prod_{t'=1}^{t-1} v(t') . \quad (51)$$

Comparison of this result to that of the steady-state analysis (Eq. (34)) reveals that as $t \rightarrow \infty$ the last term on the rhs is just A_∞ , the amplitude of the δ -function. Finally, defining $\gamma(t) = \prod_{t'=1}^t v(t')$, we find

$$(1 - \theta) = Q \sum_{j=1}^t \left(\frac{1}{j} \right)^{N/2} \frac{\gamma(t)}{\gamma(t - j)} + (1 - \theta) \left(\frac{1}{t + 1} \right)^{N/2} \gamma(t) . \quad (52)$$

Recall that at long times $v(t)$ approaches a limit $v \leq 1$. Therefore, for large t , $\gamma(t) = C_\gamma v^t$, where C_γ is a constant, up to corrections that vanish as $t \rightarrow \infty$, and

$$(1 - \theta) = Q \sum_{j=1}^t \left(\frac{1}{j} \right)^{N/2} v^j + (1 - \theta) \left(\frac{1}{t + 1} \right)^{N/2} C_\gamma v^t . \quad (53)$$

When v is strictly less than unity, the convergence is exponential, as expected from the qualitative argument given above and in agreement with the result from the

continuous-time equations below. When $\lim_{t \rightarrow \infty} v(t) = 1$, then, since the second term on the right hand side is nonzero as $t \rightarrow \infty$, we must have $\gamma(t) = \prod_{i=1}^t v(t) = C_\delta(t+1)^{N/2}$, where C_δ is a constant. This implies $v(t) \rightarrow ((t+1)/t)^{N/2} \sim 1 + O(1/t)$. By calculating the correction to the sum because of this variation in $v(t)$, we find that the amplitude of the δ -function obeys $A_t(t) - A_\infty \propto t^{-1}$ as $t \rightarrow \infty$. When $N = 2$, $v(t \rightarrow \infty)$ is exponentially close to unity, so that the decay of the amplitude in the unmutated peak up to extremely long times is governed by the logarithmic divergence of the sum for $v = 1$. Thus, in this case we expect $A_t(t)$ to decay logarithmically with t .

In figure 2 we plot $A_t(t)$, the fraction of the population which is unmutated, versus time t for the parameter values $\theta = 0.2$ and $Q = 0.1$, for $N = 1, 2$, and 3 . The curves were obtained by numerical iteration of the recursion relations Eqs. (45–48), starting with the initial condition $A_0(t = 0) = 1$ and $A_i(t = 0) = 0$ for $i \neq 0$. We find, as expected, that $A_t(t)$ decays to zero for $N = 1$ and $N = 2$ but not for $N = 3$. The scales on the plot were chosen to emphasize the logarithmic decay when $N = 2$.

To compute the probability function $\phi(\vec{x}, t)$ we first calculate the amplitudes $A_i(t)$, using the recursion relation, and then sum the corresponding Gaussians (see Eqs. (19) and (10)) to obtain numerical values of the probability. These results are plotted in figures 3, 4, and 5 for the parameter-values $\theta = 0.2$ and $Q = 0.1$. For $N = 1$, see figure 3, we plot $\phi(\vec{x}, t)$ against the magnitude of x for various values of the time, t . The picture shows a probability distribution which first gets narrower, but then settles down to a time-independent behavior for the largest times. Also drawn on this plot is the infinite-time probability, derived from Eqs. (34) and (35). The probability $\phi(\vec{x}, t)$ settles down to this limiting behavior most slowly near $\vec{x} = 0$, where fitnesses change slowly with \vec{x} .

For $N = 3$, a corresponding calculation shows a probability function which gets more and more peaked as time goes on, see figure 4. Notice how, for $t = 1000$, the peak sticks up sharply from a more slowly varying background. To see how this peaking occurs, we plot in figure 5 a scaled version of figure 4. In this version, we plot $\phi_s(\vec{x}_s, t)$ where $\phi_s(\vec{x}, t)$ is the result of dividing $\phi(\vec{x}, t)$ by the predicted growth of the peak, proportional to $t^{N/2}$, giving us a scaled y-variable $\phi_s(\vec{x}, t) = \phi(\vec{x}, t)t^{-N/2}$. The scaled x -variable is $\vec{x}_s = \vec{x}t^{1/2}$. The resulting plot becomes time-independent for large times and not-too-large values of \vec{x}_s . In this way, we see how the peak continually gets narrower, and more and more dominates the small- x behavior.

Now we turn to the question of the role of initial conditions in the behavior of discrete-time model. This question is important because the method of solution of the discrete-time model outlined in this section assumes that the initial distribution $\phi(\vec{x}, t = 0)$ is Gaussian. Here we present our arguments why the results of the

calculation should not depend sensitively on this choice of initial condition.

First, note that any initial distribution $\phi(\vec{x}, t = 0)$ which has a width much greater than both V_s and m^2 will narrow to two Gaussians of widths V_s and m^2 after one mutation step. Therefore, the behavior of any broad distribution will be indistinguishable from that of a Gaussian of width $\gtrsim m^2$.

Now we consider more general variation of the initial conditions by examining the time-dependent Eq. (51). Every term in the sum arises from a mutation event at some time $t \geq 0$; changing the initial conditions changes only the last term on the right hand side. In the $t \rightarrow \infty$ limit, the the amplitude of the δ -function is obtained by subtracting from unity the fraction of the population which has mutated at some time. The long-time limit, v , does not depend on the initial condition; to see this, note that either $v = 1$, in which case A_∞ is obtained by the subtraction just described, or else $v < 1$, in which case $A_\infty = 0$ and the amplitudes of all the terms that depend on the initial conditions vanish exponentially. Therefore, the initial conditions do not affect either the presence or absence of the δ -function or its steady-state amplitude when it is present.

4 Continuous-Time Equations

In this section we consider the continuous-time evolution described by Eq. (9). In (Bürger & Bomze, 1996) it is shown that there is a unique time-independent solution as $t \rightarrow \infty$ and the possible solutions are classified using general properties of linear operators (Reed & Simon, 1972). Our results for the long-time limit are consistent with theirs. In addition, our approach enables us to interpret simply the emergence (or not) of the δ -function peak as well to address the time-dependence of the approach to the $t \rightarrow \infty$ limit.

We proceed by Fourier transforming Eq. (9). We define the quantity $n(\vec{k}, t) = \int d^N x \phi(\vec{x}, t) \exp[i\vec{k} \cdot \vec{x}]$ and obtain:

$$\frac{\partial n(\vec{k}, t)}{\partial t} = \left[a(t) + \frac{1}{2V} \nabla_{\vec{k}}^2 \right] n(\vec{k}, t) - \Theta[1 - f(\vec{k})] n(\vec{k}, t) , \quad (54)$$

where $f(\vec{k})$ is the Fourier transform of Eq. (5):

$$f(\vec{k}) = \exp[-m^2 k^2 / 2] . \quad (55)$$

We must also specify, in addition to the initial conditions, two boundary conditions. One boundary condition is determined by the normalization requirement $\int d^N x \phi(\vec{x}, t) = 1$, which yields:

$$n(\vec{k} = 0, t) = 1 . \quad (56)$$

The second boundary condition is that $n(\vec{k}, t) \leq 1$ for all \vec{k} , which in particular implies that it cannot diverge as $\vec{k} \rightarrow \infty$. This follows from combining the normalization condition with the non-negativity requirement $\phi(\vec{x}, t) \geq 0$:

$$n(\vec{k}, t) = \int d^N x e^{i\vec{k} \cdot \vec{x}} \phi(\vec{x}, t) \leq \int d^N x \phi(\vec{x}, t) = 1. \quad (57)$$

We now solve this equation in the limit of long times. Since as $t \rightarrow \infty$ both n and a become independent of time, we must solve:

$$\left\{ -\frac{1}{2} \nabla_{\vec{k}}^2 - (\Theta V) f(\vec{k}) \right\} \psi(\vec{k}) = E \psi(\vec{k}), \quad (58)$$

where $\psi(\vec{k}) \equiv \lim_{t \rightarrow \infty} n(\vec{k}, t)$ and the eigenvalue $E = V(\lim_{t \rightarrow \infty} a(t) - \Theta)$. Eq. (58) is just the time-independent Schrödinger equation (Schiff, 1968). The parameter N , the number of characters affected by each mutation, here is interpreted as the spatial dimensionality. As shown in (Bürger & Bomze, 1996, and references therein) and as discussed here in the Appendix, the long-time solution to the population biology model is given by the lowest energy eigenstate of this Schrödinger equation.

The kinetic energy term $-\frac{1}{2} \nabla_{\vec{k}}^2$ in Eq. (58) comes from the fitness selection. This term, which in \vec{x} -space causes the distribution to become narrower and narrower, acts to cause it to become wider and wider in \vec{k} -space. The potential energy term $-\Theta V f(\vec{k})$ comes from the mutations. This potential is constant at large k , and has an attractive piece at small k . If we assume that the potential has a characteristic scale in k , so that $f(\vec{k}) = F(m\vec{k})$ (clearly the fitness function considered here is of this form), then we can define $\vec{z} = m\vec{k}$ and write the Schrödinger equation in dimensionless form:

$$\left\{ -\frac{1}{2} \nabla_{\vec{z}}^2 - \left(\frac{\Theta V}{m^2} \right) F(\vec{z}) \right\} \psi(\vec{z}) = \tilde{E} \psi(\vec{z}), \quad (59)$$

where $\tilde{E} = E/m^2$. This equation makes it clear that the nature of the behavior depends only on the single parameter $(\Theta V/m^2)$.

If only the selection (kinetic energy) term were present, the solutions to Eq. (58) consistent with the normalization condition $\phi(\vec{k}) = 1$ are

$$\psi(\vec{z}) = \exp[i\vec{p} \cdot \vec{z}], \quad (60)$$

where $\vec{p} = \hat{z} \sqrt{2\tilde{E}}$, and \hat{z} is a unit vector along \vec{z} . To be consistent with the boundary conditions, we require $\tilde{E} \geq 0$. The ground state eigenstate thus has $\tilde{E} = 0$, so that $\psi(\vec{z}) = 1$. Fourier transforming this result, we see that in the absence of mutations, at $t = \infty$ the entire population has the same genotype. This result is consistent with our calculation in the previous section demonstrating that in the absence of mutations the variance of $\phi(\vec{x}, t)$ narrows in proportion to $1/t$.

So now we consider the effects of adding the potential arising from the mutation term. This potential is attractive, so that the question we must address is whether the ground state eigenstate remains extended ($\psi(\vec{z}) \rightarrow \text{constant}$ as $|\vec{k}| \rightarrow \infty$) or whether it results in a bound state with $\tilde{E} < 0$, which has $\psi(\vec{z}) \rightarrow 0$ exponentially in k as $\vec{k} \rightarrow \infty$. In the latter case, the real-space distribution $\phi(\vec{x})$ is smooth as $\vec{x} \rightarrow 0$. Therefore, if there is a bound state, then only an infinitesimal fraction of the population is unmutated, whereas if there are no bound states, then a finite fraction of the population has a unique genotype.

It has been proven for the Schrödinger equation that any attractive potential, no matter how small, will lead to a bound state for $N \leq 2$ (Simon, 1976). Thus, the population distribution in real space $\phi(\vec{x}, t \rightarrow \infty)$ must be smooth for $N \leq 2$. When $N > 2$ bound states only appear if the potential is large enough ($\Theta V_s/m^2 > C_N$, where C_N is of order unity) (Schiff, 1968). Therefore, in this case, if the potential is small (weak mutation effects), then the ground state remains extended: $\psi(\vec{k}) \rightarrow \text{constant}$ as $\vec{k} \rightarrow \infty$, and the real space distribution $\phi(\vec{x})$ has a δ -function piece. If the potential is large (strong mutation effects), then there is a bound state, and the distribution $\phi(\vec{x})$ is not singular at small \vec{x} .

If all the states are extended, then as $|\vec{k}| \rightarrow \infty$ the lowest energy state obeys $\nabla_{\vec{k}}^2 n(\vec{k}) = 0$ and has the form $n(\vec{k}) \rightarrow A + Bk^{-(N-2)}$, where A and B are coefficients determined by matching to the solution in the small $|\vec{k}|$ region.¹¹

For almost all potentials both A and B are nonzero; Fourier transforming the term proportional to B yields:

$$\text{B term} \propto \int d\Omega \int_0^\infty dk k^{N-1} e^{ikx \cos(\theta)} k^{-(N-2)} \propto x^{-2}. \quad (61)$$

(Here, $\int d\Omega$ is the integral over angles.) Thus, just as in our discrete-time solution, we find associated with a δ -function at $\vec{x} = 0$ a power-law divergence at small x , $\phi(\vec{x}) \propto x^{-2}$.

In the regime where the ground state is bound, there is a finite energy gap between the ground state and the excited states, which implies that the long-time limit is approached exponentially in time. When there is no bound state, the energy spectrum of the Schrödinger equation is continuous, which leads to power-law convergence to the long-time limit. These points are discussed in greater detail in the Appendix.

¹¹This form can be verified for all N by several means, the most straightforward being direct evaluation using Cartesian coordinates. A more elegant method is to note that we are looking for a function g which satisfies $0 = \nabla^2 g = \text{div} \cdot \text{grad } g$. Since $\text{grad } g$ is a current \vec{j} that is angle-independent and obeys $\text{div} \vec{j} = 0$, then $\int j dS_r = 0$, where S is the surface of a sphere of radius r . Therefore, one must have $j \propto r^{-(N-1)}$, so that $g \propto r^{-(N-2)}$.

5 Relation between the continuous-time and the discrete-time models

In this section we discuss the similarities and differences between the results of our discrete-time analysis valid when $V_s/m^2 \ll 1$ and those of the continuous-time analysis. The basis of our discussion is Eq. (8), which can be used to describe both cases.

Qualitatively, the discrete and continuous models exhibit very similar behavior. They both lead to only smooth distributions (a “normal” phase) for $N \leq 2$, and for $N > 2$ display a transition as the mutation rate is decreased between the normal phase and a “condensed” phase, where a finite fraction of the population has a unique genotype. In both models, the convergence to the long-time limit is power law in the condensed phase and is exponential in the normal phase. However, there are quantitative differences between the two models, particularly the detailed dependence on model parameters.

In the continuous-time model, the form of Eq. (59) makes it clear that the behavior depends only on the single parameter $\Theta V/m^2$. In contrast, the discrete model that we solve in the limit $m^2/V_s \gg 1$ (Eqs. (24-26)) depends on two parameters: $Q = (1 + m^2/V_s)^{-N/2}$ and θ . For $N > 2$, the discrete model might or might not have a condensate, depending on the value of the ratio $\theta Q/(1 - \theta)$. In this regime, the continuous-time model is always in the condensed phase.

We use Eq. (8) to discuss the relation between the continuous- and discrete-time models. One can Fourier transform this equation (as we did in section 4 for the continuous-time model) and interpret the result as a quantum-mechanical problem of a particle in a potential which is applied only at the discrete times $t = n\tau$ as in Eq. (8).¹²

The discrete model has two parameters while the continuous model has only one, since Eq. (8) depends not just on the mean mutation rate Θ but also on τ , the time between mutation events. Thus, in principle the discrete-time model has one more parameter than the continuous-time one. If τ is increased by a factor of two, then the product θV (θ the mutation probability per generation and V the viability selection parameter) remains constant, but V itself is halved. The analysis in section 3, which is valid when V/m^2 is very small, and hence in the limit when τ is very long, yields a solution in which that both θ and V/m^2 enter explicitly and not just as a product. Nonetheless, two models have very similar outcomes.

¹²We take the limit where a very strong potential is applied for a very short time at each $t_n = n\tau$.

6 Comparison with sequence database statistics

Here we present our investigations of the statistical properties of DNA sequences archived in genetic databases. The aim is to see if observed sequence variability of a gene is correlated with its degree of pleiotropy. The focus in this section is on a qualitative prediction of the model, that genes with a high degree of pleiotropy should have a narrower distribution of alleles than those that affect only one trait. This trend, which is consistent with the mathematical results of the previous sections, is easy to understand: if a gene of high fitness is pleiotropic, then because each mutation affects several characters independently, the chances are high that a given mutation leads to a large fitness decrease.

Our first test for possible correlation between degree of pleiotropy and the probability distribution describing genetic variation is simply to count the number of naturally occurring variants of various genes in the *Drosophila* database FlyBase (FlyBase, 1998). We choose to include naturally occurring alleles, including spontaneous mutations arising in laboratory stocks, of the genes analyzed. Mutagen-induced variants are not included, since there is no evidence that they exist in natural populations. Table 1 shows data for nine genes: *brown*, *cinnabar*, *ecdysone receptor*, *engrailed*, *fork head*, *hairless*, *notch*, *vestigial*, and *white*. The number of naturally occurring variants is tabulated together with the length of the primary transcript and the approximate length of the genomic region encompassing the gene and its flanking regulatory regions. Transcript and genomic lengths are taken from the full-format gene reports in FlyBase. For genes with multiple transcripts, we report the length of the longest described transcript. When the full gene report does not include the genomic length, this is estimated from the FlyBase molecular map of the gene. The number of variants per gene is normalized both by the transcript length and by the genomic length. In both cases, the values range over approximately three orders of magnitude. We estimate the degree of pleiotropy of each of the listed genes according to the number of tissues/structures and developmental stages in which the genes are expressed. This ordering is admittedly arbitrary to some extent, but few would argue with the conclusion that *fork head* (encoding a transcription factor essential for development of the midgut) and *engrailed* (encoding a homeotic gene establishing segmentation) are more pleiotropic than the eye color mutants *cinnabar*, *white*, and *brown*. Inspection of Table 1 reveals a clear inverse correlation between estimated pleiotropy and number of variants/transcribed length or variants/genomic length. This relationship is consistent with the prediction of Waxman and Peck's model.

A weak point of this approach is that other factors besides degree of pleiotropy could lead to the observed differences in the number of variants. For instance, even

two genes which only affect single traits could have vastly different variabilities, if the traits have greatly different effects on overall fitness. For example, a gene that controls eye color may well have a much smaller overall effect on fitness than a gene that controls a crucial developmental function. In the population genetics model, differences of this type are reflected by different values of the fitness parameter V_s . Within the model, differences in overall variability caused by changing V_s can be distinguished from those caused by changing the degree of pleiotropy N because they yield very different functional forms for the distribution function $\phi(\vec{x})$. However, counting mutations yields no information about the functional form of the distribution function and thus cannot be used to distinguish between the mechanisms.

Our second strategy for relating degree of pleiotropy to statistics of archived DNA sequences aims to obtain information about the form of the probability distribution for various genes. It assumes that two genetic variants whose sequences are highly similar in sequence space code for genes that are close together in fitness space. We use the BLAST 2 sequence similarity search tool (Altschul *et al.*, 1997) to search the GENBANK and EMBL DNA sequence databases¹³ against the 8 gene sequences listed in Table 2. These databases contain sequences of genes from many organisms, though a preponderance of the sequences are from humans. Given a target sequence, BLAST 2 generates a list of matching sequences along with scores which measure the degree of similarity. Because the maximum possible score for a given sequence is larger for longer sequences, we avoided examining genes that were either very short or very long. As can be seen in table 2, the lengths of the sequences examined varied by less than a factor of ten. Only genes were used whose list of matches was truncated because the score fell below the cutoff value of 40; no genes were used whose match list was truncated because the number of matching sequence exceeded BLAST 2's maximum number of matches of 500. As shown in table 2, the number of matches obtained also varies by roughly an order of magnitude for the genes in the table, and is not obviously correlated with the length of the gene.

High BLAST 2 scores correspond to sequences which match the search sequence the most closely. The scores range from roughly 10^4 (a sequence matching itself) to the default cutoff of 40. We assume that this score is inversely correlated with the evolutionary distance between the input gene and the retrieved sequences, so that a gene with a high fraction of matches with low scores has a broader probability distribution $\phi(x)$ than a gene with more matches with high scores. Thus, if all the matches tend to be very good, the gene is considered to be less variable than one with many poor matches. We examine the statistics of the closeness of the matches,

¹³These databases and the BLAST 2 search tool are maintained by the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>).

in particular the fractions of the total number of matches that fall into given ranges of scores.

Figure 6 is a plot of a histogram of the fraction of the matches in a given score range obtained versus the inverse of the score for several genes. We do observe significant variability between these genes; for example, *cytochrome P-1-450*, which codes for a general-purpose antioxidant expressed in many tissues and is thus plausibly highly pleiotropic, has relatively many very good matches compared to *rhodopsin*, which is quite specific, coding for a protein essential for vision. However, we cannot claim that there is an unambiguous correlation between degree of pleiotropy and these plots. For example, *HOX A1* has many poor matches, where it might be expected to be pleiotropic, since it is essential to many developmental functions. Thus, though interesting variations in the statistics of DNA sequences for different genes are found, we have not been able to demonstrate convincingly a correlation between degree of pleiotropy and these statistical variations.

The behavior of *HOX A1* can be explained by a process of repeated duplication and divergence over the course of evolution. Many of the retrieved genes have assumed distinct but related developmental functions. The BLAST 2 analysis presented here does not account for a shift in score distributions arising from divergence among members of gene families and superfamilies. Another fundamental difficulty with our analysis is illustrated by the case of *alpha-tubulin*, which codes for a protein vital to the formation of microtubules. On the one hand, microtubules are expressed in many different tissues, but on the other hand, all its functions arise from similar structural properties. Thus, one could categorize *alpha-tubulin* as either highly pleiotropic (since it is expressed in many tissues) or as non-pleiotropic (since the function is similar everywhere where it is expressed). In short, though degree of pleiotropy is defined in the populations genetic model, it is not clear how to quantify it in terms of biological function.

There are several additional difficulties and ambiguities inherent in comparisons between genome sequence data and population genetics models of the type considered here. First, the database scores are based on the quality of sequence alignments, whereas the population genetics models define distances in terms of fitness, which is a phenotypic quantity. Fitness distance and evolutionary sequence distance may be quite different; for example, point mutations at different locations may range in effect from unobservable to lethal. However, even if there is not a single definite relationship between sequence distance and fitness distance, so long as these quantities are positively correlated with each other, meaningful results may be obtained. Variability in the relation between the distances will result in noisy data, but different genes are all subject to similar uncertainties. Therefore, it is plausible to hope that, if sig-

nificant differences between the sequence variability of two genes are observed, they reflect differences in the fitness distribution $\phi(\vec{x})$ of the population genetics models. Another limitation of our analysis is that we consider the fitnesses of alleles as fixed. This is not necessarily the case, as many gene products function as components of multi-protein structures or pathways (Lewontin, 1974), (Dover & Flavell, 1984). This dependency of the fitness value of a specific sequence variant on the remainder of the genotype is manifested explicitly by the existence of epistatic interactions (e.g., (Fijneman *et al.*, 1996), (van Wezel *et al.*, 1996)).

In addition, one may worry that the population genetics model we have used involves a continuously varying phenotype, whereas sequence variations are discrete. However, because genes are thousands of base pairs long, there is still a huge range of variability, and the use of a continuum model is therefore reasonable.

Finally, as mentioned above, pleiotropy itself is defined in terms of phenotypic fitness, and it is not clear how to create an independent measure that enables one to assign objectively the degree of pleiotropy N to a given gene. This is a serious fundamental difficulty that must be overcome if one is to make meaningful quantitative statistical analysis of the possible correlations between the sequence statistics and the degree of pleiotropy.

7 Discussion

In this paper we have analyzed some variants of a population biology model incorporating selection, mutation, and pleiotropy. We have focussed on understanding the circumstances under which at long times a nonzero fraction of the population has a unique genotype, and on characterizing the time dependence of the population distribution in this regime. We have analyzed both the discrete-time model of (Waxman & Peck, 1998) as well as an associated continuous-time model. We find:

1. In both the discrete and continuous-time models, a unique genotype can emerge only when N , the number of characters affected by each gene, is greater than two, a result in agreement with (Waxman & Peck, 1998). For any $N > 2$, the infinite-time population distribution $\phi(\vec{x}, t \rightarrow \infty)$ contains a δ -function contribution when the mutation rate is nonzero but small, but not when the mutation rate is large enough.
2. In the regime where $\phi(\vec{x}, t \rightarrow \infty)$ has a δ -function contribution, this δ -function is accompanied by a $1/x^2$ singularity at small \vec{x} .
3. The δ -function peak emerges as the limit of a peak that continually becomes

higher and narrower. Thus, in this regime the convergence to the $t \rightarrow \infty$ limit is power-law.

4. In the regime when $\phi(\vec{x}, t \rightarrow \infty)$ is smooth, the convergence to this distribution is exponential in time.
5. The continuous- and discrete-time models exhibit qualitatively identical behavior, but there are quantitative differences between them.

Our analysis of the discrete-time equations relies on the use of Gaussian functions for both the fitness and mutation terms. However, our analysis of the continuous-time model assumes only that the potential in the quantum mechanics problem is attractive and short-ranged. Therefore, we expect our results to apply qualitatively for a large class of different mutation terms.

Other mechanisms giving rise to genetic diversity such as antagonistic pleiotropy, genetic drift, and discreteness of alleles change qualitatively the nature of the equations describing the system. It will be interesting to see whether the “condensation” phenomenon investigated here is robust when these effects are taken into account.

In addition to our mathematical analysis, we also attempted to assess whether the degree of pleiotropy of genes resulted in systematic trends in the degree of variation they exhibited in the mutations documented in online genetic databases. These attempts to correlate the degree of pleiotropy with the statistics of the sequence variations were inconclusive.

Available database resources do not allow unbiased sampling of the sequence variation present in a natural population. Such a database is likely to emerge as the Human Genome Project’s present initiative to sample the extent of sequence variation in a wide array of genes in the American population progresses.

8 Acknowledgments

We are grateful for support by the National Science Foundation, Grant DMR 96-26119 (SNC), and the Office of Naval Research, Grant N00014-96-1-0127 (LPK). SNC and RDB thank J.J. Sohn for useful suggestions.

Appendix: The relation between eigenstates of the Schrödinger equation and the long-time behavior of the continuous-time population genetics model

In section 4 above we use the result that the long-time behavior of Eq. (54) is given by the lowest energy eigenstate of the associated Schrödinger Eq. (58). This result has been shown by Bürger and collaborators (Bürger & Bomze, 1996, and references therein); for completeness we derive it here and emphasize the implications for the approach to the long-time behavior in the population biology model.

We assume that we know the solution of the time-independent problem, which is a set of eigenstates $u_n(\vec{z})$ which satisfy:

$$-\frac{1}{2}\nabla_z^2 u_n(\vec{z}) + V(\vec{z})u_n(\vec{z}) = E_n u_n(\vec{z}) . \quad (62)$$

The eigenstates $u_n(\vec{z})$ are complete (any function can be written as a linear combination of them) and can be chosen to be orthonormal:

$$\int d^N z u_{n_1}^*(\vec{z}) u_{n_2}(\vec{z}) = \delta_{n_1, n_2} , \quad (63)$$

where the star denotes complex conjugate. This normalization is usual in quantum mechanics (Schiff, 1968).

We need to examine the time-dependent Eq. (54); using scaled variables, it is written as:

$$-\frac{\partial \psi(\vec{z}, t)}{\partial t} = -\left[a(t) + \frac{1}{2}\nabla_z^2\right] \psi(\vec{z}, t) + V(\vec{z})\psi(\vec{z}, t) , \quad (64)$$

where

$$a(t) = \left[-\frac{1}{2}\nabla_z^2 + V(z)\right] \psi(\vec{z}, t)|_{\vec{z}=0} . \quad (65)$$

We write the wave function at time t_0 , $\psi(\vec{z}, t_0)$, as a linear combination of eigenstates:

$$\psi(\vec{z}, t_0) = \sum_i b_i(t_0) u_i(\vec{z}) . \quad (66)$$

The condition $\psi(\vec{z} = 0, t_0) = 1$ is enforced by requiring $\sum_i b_i(t_0) u_i(\vec{z} = 0) = 1$. By substituting the form (66) for ψ and using the orthonormality of the eigenstates, one finds equations for the time-evolution of the coefficients b_i :

$$-\frac{db_i(t)}{dt} = b_i(t) [E_i - a(t)] . \quad (67)$$

Integrating this equation yields:

$$b_i(t) = b_i(t_0) \mathcal{A}(t) e^{-E_i t} , \quad (68)$$

where $\mathcal{A}(t) = \exp \left[\int_{t_0}^t ds a(s) \right]$. Thus the ratio of the weights of any two eigenstates i_1 and i_2 is given by:

$$\frac{b_{i_1}(t)}{b_{i_2}(t)} = \frac{b_{i_1}(t_0)}{b_{i_2}(t_0)} \exp [-(E_{i_1} - E_{i_2})t] . \quad (69)$$

At long times the exponential factor ensures that the state with the largest weight is the one with the lowest energy.

If there is a nonzero energy gap between the lowest and second-lowest energy eigenstates (no δ -function in $\phi(\vec{x})$), then Eq. (69) implies that the approach to the long-time limit is exponential in time. If there is no bound state, then all the states are extended and the spectrum is continuous; the E_i take the form $E_i = q_i^2/2$, where q_i has the dimension of a wavevector in \vec{z} space, (Schiff, 1968) and hence a distance in the original \vec{x} space. Therefore,

$$\psi(\vec{z}, t) = \mathcal{A}(t) \sum_i b_i(t_0) e^{-q_i^2 t/2} . \quad (70)$$

The relative weight of the eigenstate i remains large until $q_i \sim \sqrt{1/t}$. Thus, the δ -function at infinite times emerges from a peak that is narrowing, having a width proportional to $\sqrt{1/t}$.

References

- Abramowitz, M. and Stegun, C. A. (Eds.) (1972). Riemann Zeta Function and Other Sums of Reciprocal Powers. Chapter 23.2 in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th printing. New York: Dover, pp. 807-808.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402; <http://www.ncbi.nlm.nih.gov/BLAST/>
- Barton, N.H. & Turelli, M. (1989). Evolutionary quantitative genetics: how little do we know? *Annu. Rev. Genet.* **23**, 337-370.
- Bürger, R. (1986). On the maintenance of genetic variation: global analysis of Kimura's continuum-of-alleles model. *J. Math. Biol.* **2**, 341-351.
- Bürger, R. (1986). Mutation-selection balance and continuum-of-alleles models. *Mathematical Biosciences* **91**, 67-83.

- Bürger, R. and Hofbauer, J. (1994). Mutation load and mutation-selection-balance in quantitative genetic traits. *J. Math. Biol.* **32**, 193-218.
- Bürger, R. & Bomze, I.M. (1996). Stationary distributions under mutation-selection balance: Structure and properties. *Adv. Appl. Prob.* **28**, 227-251.
- Carrier, G.F., Krook, M., Pearson, C.E. (1983). *Functions of a Complex Variable*. Hod books, Ithaca, N.Y. p.192 ff.
- de Groot, S.R., Hooyman, G.Y., ten Seldam, C.A. (1950). On the Bose-Einstein Condensation. *Proc. Roy. Soc. A* **203**, 266-286.
- Dover, G.A. & Flavell, R.B. (1984). Molecular coevolution: DNA divergence and the maintenance of function. *Cell* **38**, 622-623.
- Feynman, R.P. (1972). *Statistical Mechanics: A Set of Lectures*, Benjamin/Cummings, Reading, MA), pp. 30-34.
- Fijneman, R.J., de Vries, S.S., Jansen, R.C., Demant, P. (1996). Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse. *Nat. Genet.* **14**, 465-467.
- Fröhlich, J. & Pfister, C. (1981). On the absence of spontaneous symmetry breaking and of crystalline order in two-dimensional systems. *Commun. Math. Phys.* **81**, 277-298.
- The FlyBase Consortium (1998). FlyBase – A *Drosophila* Database. *Nucleic Acids Research* **26**, 85-88; <http://flybase.bio.indiana.edu/>.
- Gavrilets, S. & de Jong, G. (1993). Pleiotropic models of polygenic variation, stabilizing selection, and epistasis. *Genetics* **134**, 609-625 (1993).
- Hohenberg, P.C. (1967). Existence of long-range order in one and two dimensions. *Phys. Rev.* **158**, 383-386.
- Kimura, M. (1965). A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proc. Nat. Acad. Sci. USA* **54**, 731-736.
- Kingman, J.F.C. (1978). A simple model for the balance between selection and mutation. *J. Appl. Prob.* **15**, 1-12.
- Lande, R. (1980). The genetic covariance between characters maintained by pleiotropic mutation. *Genetics* **94**, 203-215.

- Lewin, L. *Polylogarithms and Associated Functions*. North-Holland, New York.
- Lewontin, R.C. (1974). *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.
- Lichtenberg, A.J. & Leiberman, M.A. (1992). *Regular and Chaotic Dynamics*. Springer Verlag Applied Mathematical Sciences series, number 38, Springer Verlag, New York, p. 219.
- Mermin, D. and Wagner, H. (1966). Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic Heisenberg models. *Phys. Rev. Lett.* **17**, 1133-1136.
- Reed, M. and Simon, B. (1972). *Methods of modern mathematical physics, Vol.4: Analysis of operators*, Academic Press, New York.
- Schiff, L.I. (1968). *Quantum Mechanics*, third edition, McGraw-Hill, New York.
- Simon, B. (1976). The Bound State of Weakly Coupled Schrödinger Operators in One and Two Dimensions. *Annals of Physics* **97**, 279-288.
- Spanier, J. & Oldham, K. B. (1987). *An Atlas of Functions*, Hemisphere Pub., Washington, DC, chapter 10.
- Turelli, M. (1984). Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoret. Pop. Biol.* **25**, 138-193.
- Turelli, M. (1985). Effects of pleiotropy on predictions concerning mutation-selection balance for polygenic traits. *Genetics* **111**, 165-195.
- van Wezel, T., Stassen, A.P., Moen, C.J., Hart, A.A., van der Valk, M.A., Demant, P. (1996). Gene interaction and single gene effects in colon tumour susceptibility in mice. *Nat. Genet.* **14**, 468-470.
- Wagner, G.P. (1989). Multivariate mutation-selection balance with constrained pleiotropic effects. *Genetics* **122**, 223-234.
- Waxman, D. & Peck, J.R. (1998). Pleiotropy and the preservation of perfection. *Science* **279**, 1210-1213.

gene	PR	N_V	L_T	L_G	N_V/L_T	N_V/L_G
<i>fork head</i>	1	1	4.0	13 - 60	2.5×10^{-4}	$(1.7 - 7.7) \times 10^{-5}$
<i>engrailed</i>	2	19	3.6	200	5.3×10^{-4}	9.5×10^{-5}
<i>ecdysone receptor</i>	3	13	5.5	35	2.4×10^{-3}	3.7×10^{-4}
<i>notch</i>	4	70	15	80 - 150	4.7×10^{-3}	$(4.7 - 8.8) \times 10^{-4}$
<i>vestigial</i>	5	55	3.8	46	0.014	1.2×10^{-3}
<i>hairless</i>	6	17	6.0	8	2.0×10^{-3}	2.1×10^{-3}
<i>cinnabar</i>	7	22	2.5	15	8.8×10^{-3}	1.5×10^{-3}
<i>white</i>	8	247	6.0	48 - 200	0.041	$(1.2 - 5.1) \times 10^{-3}$
<i>brown</i>	9	61	3.0	140	0.020	4.4×10^{-4}

Table 1: Table of number of naturally-occurring mutations for different *Drosophila* genes. PR is pleiotropy rank, estimated using the number of tissues/structures and developmental stages in which the genes are expressed. N_V is the number of naturally-occurring variants of the gene, L_T is the transcript length, and L_G is the genomic length. All lengths are given in kilobases.

gene	length	number of matches
<i>cytochrome P-1-450</i>	2565	81
<i>dystrophin</i>	13957	216
<i>epidermal growth factor receptor</i>	2660	57
<i>HOX A1</i> (human)	2595	380
<i>huntingtin</i>	10348	108
<i>iodothyronine deiodinase</i>	2222	159
<i>rhodopsin</i>	6953	208
<i>alpha-tubulin</i>	1596	488

Table 2: Genes examined using BLAST 2 sequence similarity search tool. Lengths of genes are given in units of the number of bases.

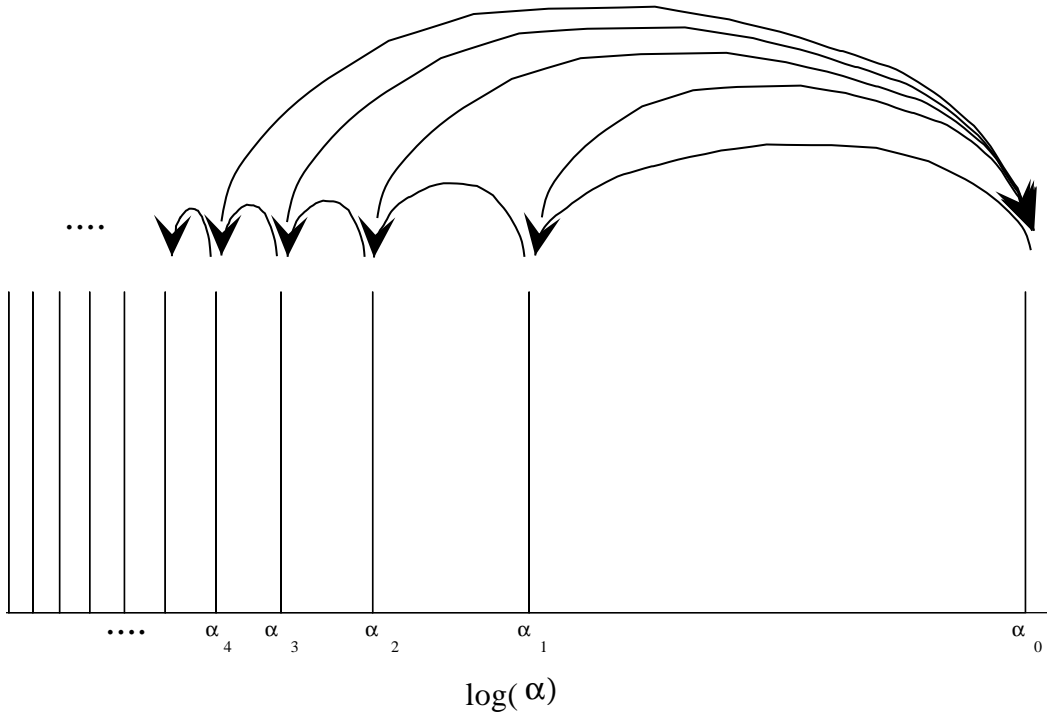


Figure 1: Sketch of transitions between α_i 's in the limit $V_s/m^2 \ll 1$. Each α_i corresponds to a Gaussian in $\phi(\vec{x}, t)$.

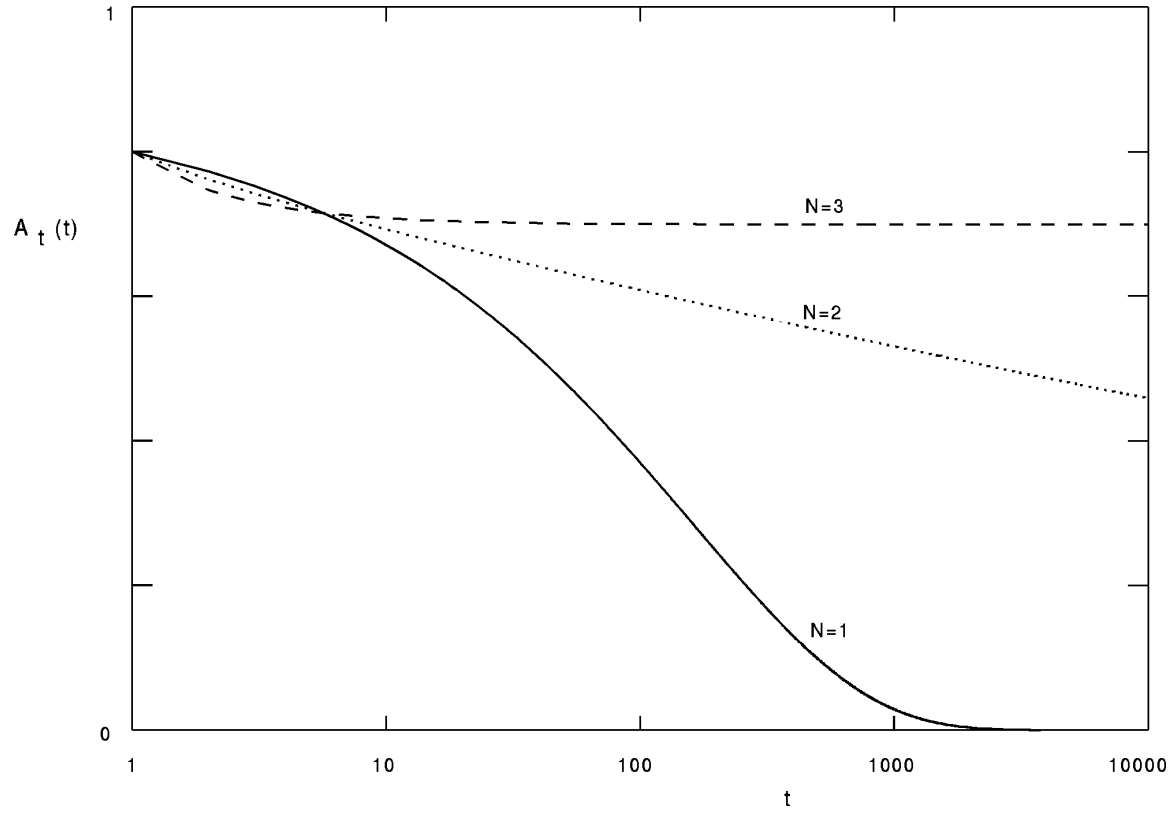


Figure 2: Plot of the fraction of the population which is unmutated, $A_t(t)$, versus time t for the discrete-time model with parameter values $Q = 0.1$, $\theta = 0.2$, obtained by numerical iteration of Eqs. (45–48) starting with $A_0(t = 0) = 1$, $A_{i \neq 0}(t = 0) = 0$. For $N = 1$ and 2, $A_t(t)$ decays so that $A_\infty \equiv \lim_{t \rightarrow \infty} A_t(t) = 0$, while for $N = 3$, A_∞ is nonzero.

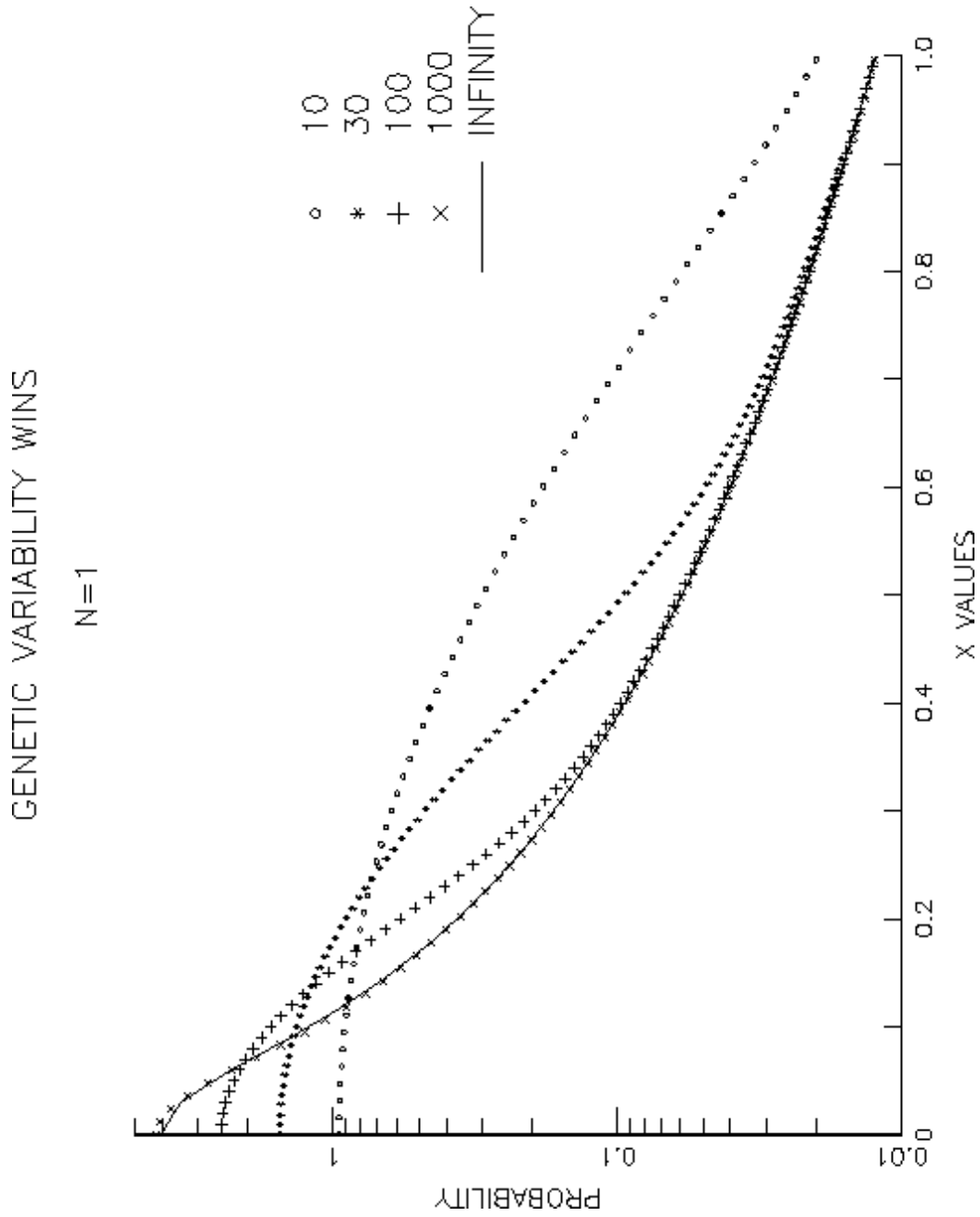


Figure 3: Plot of probability function $\phi(\vec{x}, t)$ against the magnitude of \vec{x} for $N = 1$. The numerical results for times 10, 30, 100, and 1000 are calculated by summing Gaussians with weights computed from the recursion relations. This set of curves is compared with the expected infinite-time result, calculated as a solution to Eqs. (34) and (35). For large times, the two types of calculations agree quite well.

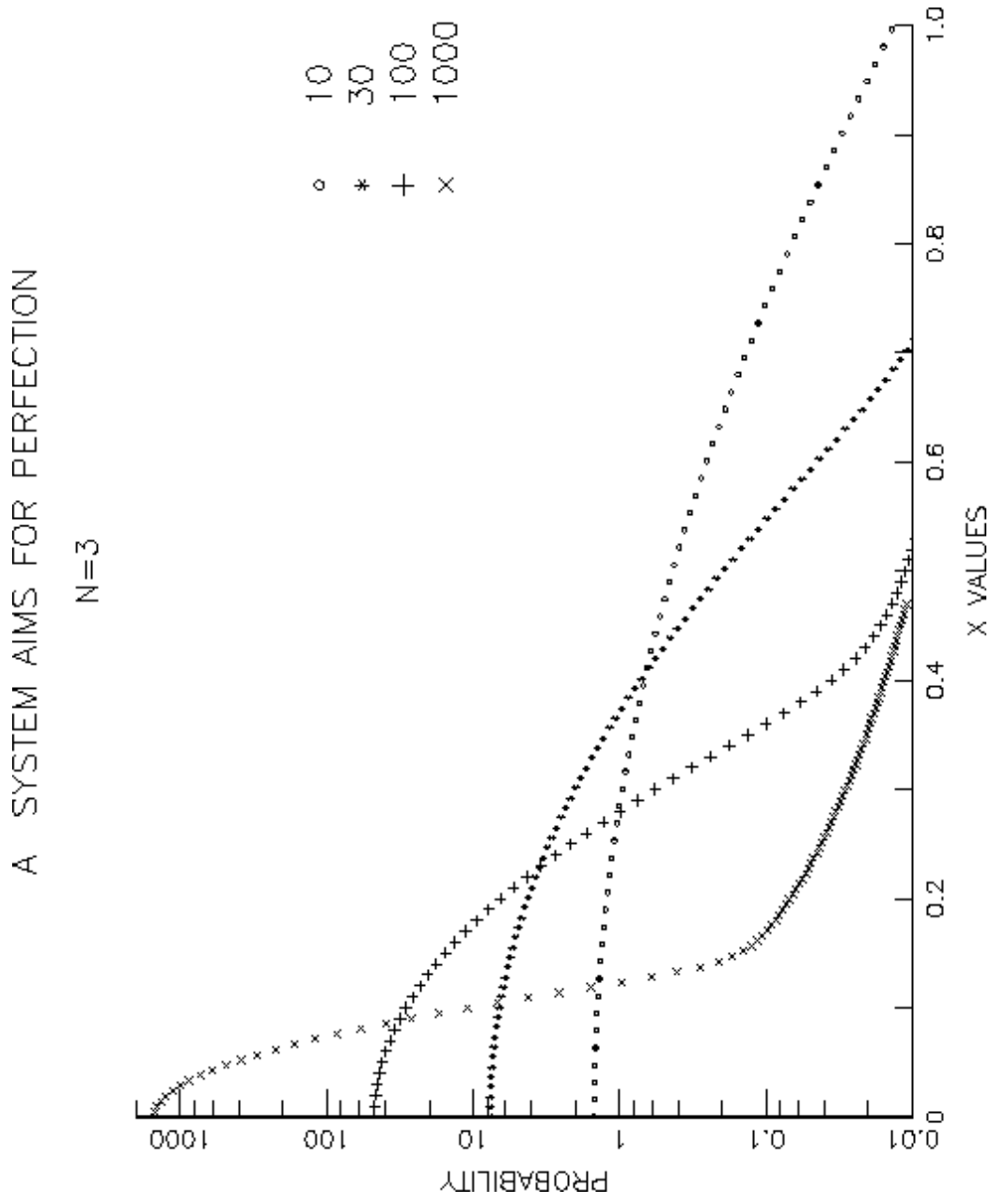


Figure 4: Plot of probability function $\phi(\vec{x}, t)$ against the magnitude of \vec{x} for $N = 3$. The numerical results are shown for times 10, 30, 100, and 1000. In contrast to the $N = 1$ case, these curves contain a central peak which continues to narrow and to grow for large t .

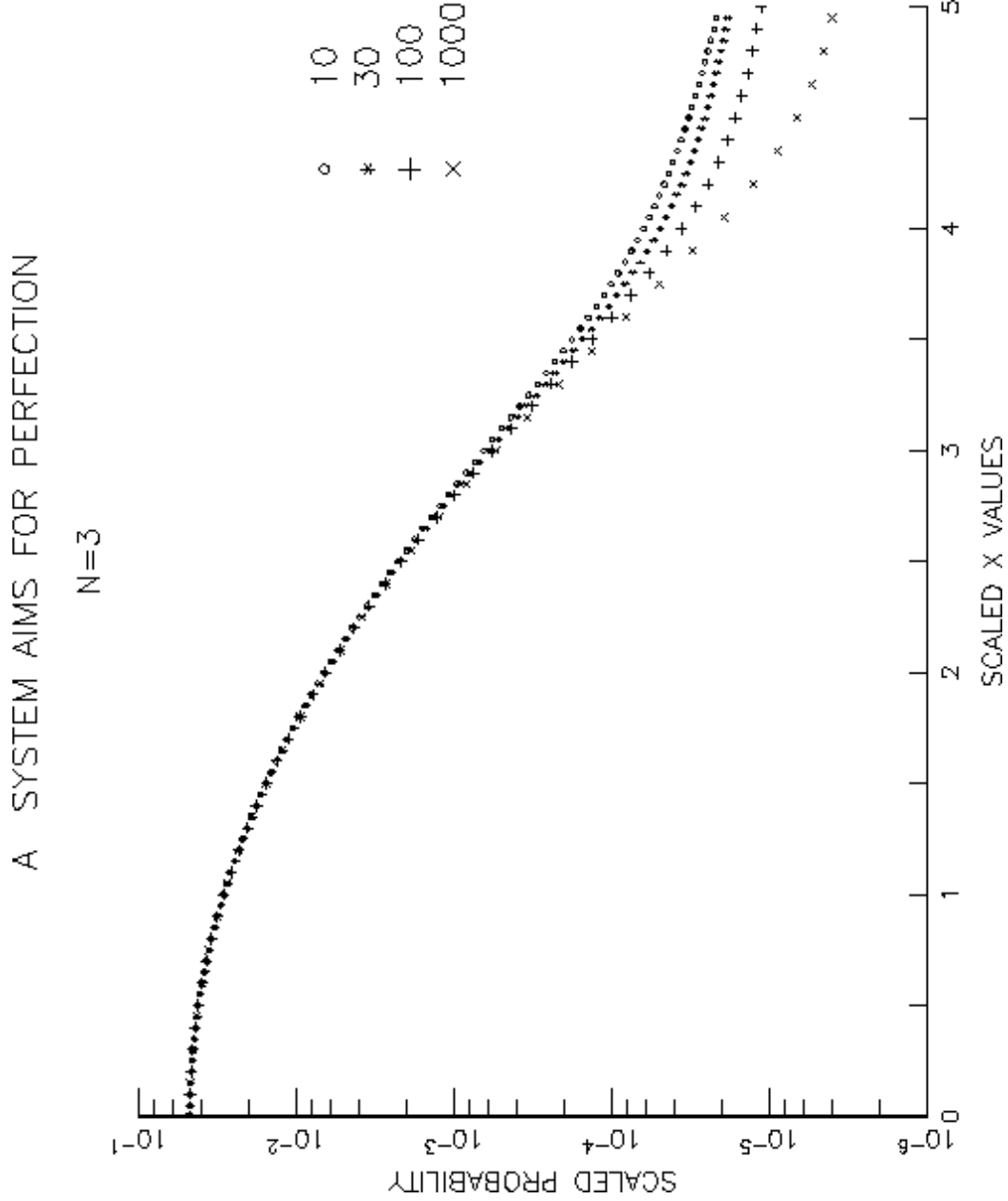


Figure 5: Scaled plot of probability function versus the magnitude of x for $N = 3$. The scaling is picked to make the central peak show a time independent behavior in the new coordinates. The abscissa is $xt^{1/2}$, so that for larger time the picture focuses upon smaller values of x . The ordinate is $\phi(\vec{x}, t)t^{-N/2}$, and thereby the picture focuses upon increasingly concentrated distributions for larger t . Numerical results are shown for times 10, 30, 100, and 1000. In these coordinates, the figure shows (as expected) an approach to a constant value at large times.

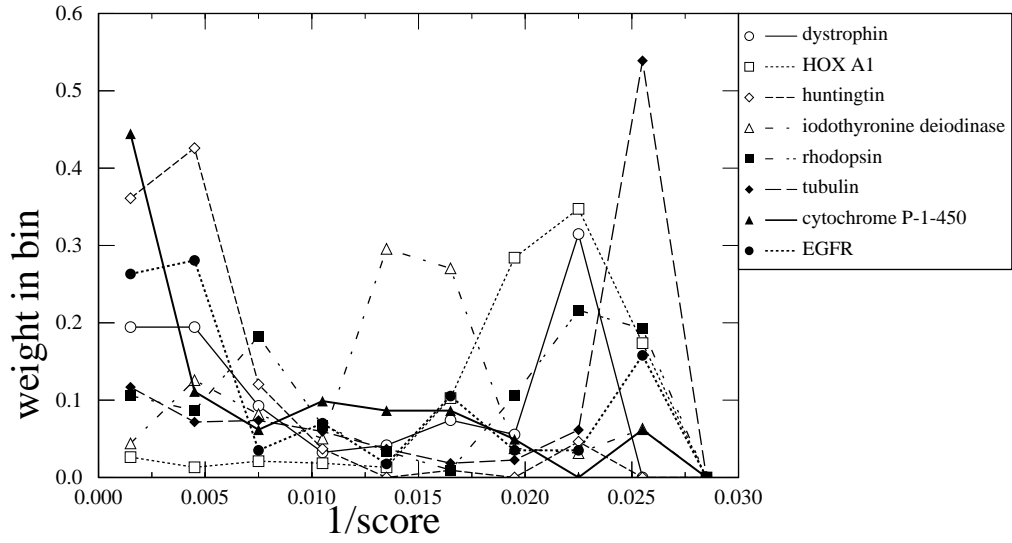


Figure 6: Histogram plot of the fraction of matches with inverse scores in a given range, as a function of inverse score, for several genes with differing degrees of pleiotropy. Significant differences in the statistics of these matches are observed; for example, *cytochrome P-1-450* has a very large percentage of matches with high scores, whereas the matches for *rhodopsin* tend to have low scores.